Linear Regression

Nipun Batra and the teaching staff

IIT Gandhinagar

July 30, 2025

$$\textit{weight}_i \approx \theta_0 + \theta_1 \cdot \textit{height}_i$$

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times d}\boldsymbol{\theta}_{d\times 1}$$

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times d}\boldsymbol{\theta}_{d\times 1}$$

- θ_0 - Bias Term/Intercept Term

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times d}\boldsymbol{\theta}_{d\times 1}$$

- θ_0 - Bias Term/Intercept Term
- θ_1 - Slope

Examples in multiple dimensions.

Examples in multiple dimensions.
One example is to predict the water demand of the IITGN campus

Examples in multiple dimensions.
One example is to predict the water demand of the IITGN campus

> Demand = f(# occupants, Temperature)

Examples in multiple dimensions.
One example is to predict the water demand of the IITGN campus

Demand = f(# occupants, Temperature)

Demand = Base Demand + K_1 * # occupants + K_2 * Temperature

$$
\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N\times(M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1)\times 1}
$$

$$\begin{bmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_N} \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N\times(M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1)\times 1}$$

$$\boxed{\hat{Y} = X\theta}$$

$$Y = X\theta + \epsilon$$

$$Y = X\theta + \epsilon$$

To Learn: θ

$$Y = X\theta + \epsilon$$

To Learn: θ
Objective: minimize $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$

Objective: Minimize $\epsilon^T\epsilon$

The above table represents the data after transformation

The above table represents the data after transformation
Now, we can write $\hat{s} = f(t, t^2)$

The above table represents the data after transformation
Now, we can write $\hat{s} = f(t, t^2)$
Other transformations: $\log(x), x_1 \times x_2$

A linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_i$ is of the following form

A linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_i$ is of the following form

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 + \cdots + \alpha_i\mathbf{v}_i$$

where $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_i \in \mathbb{R}$

The span of $v_1, v_2, \ldots, v_i$ is denoted by $\text{SPAN}\{v_1, v_2, \ldots, v_i\}$

The span of $v_1, v_2, \ldots, v_i$ is denoted by SPAN$\{v_1, v_2, \ldots, v_i\}$

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

The span of $v_1, v_2, \ldots, v_i$ is denoted by SPAN$\{v_1, v_2, \ldots, v_i\}$

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

It is the set of all vectors that can be generated by linear combinations of $v_1, v_2, \ldots, v_i$.

The span of $v_1, v_2, \ldots, v_i$ is denoted by SPAN$\{v_1, v_2, \ldots, v_i\}$

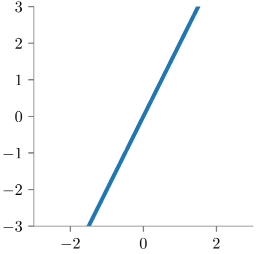
$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

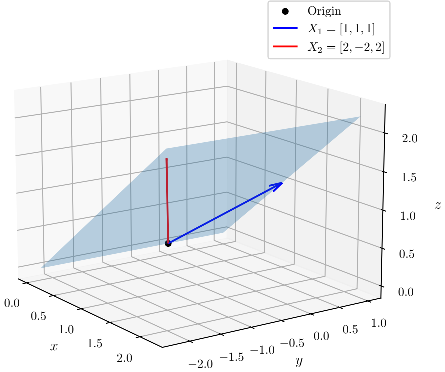
It is the set of all vectors that can be generated by linear combinations of $v_1, v_2, \ldots, v_i$.
If we stack the vectors $v_1, v_2, \ldots, v_i$ as columns of a matrix V, then the span of $v_1, v_2, \ldots, v_i$ is given as $V\alpha$ where $\alpha \in \mathbb{R}^i$

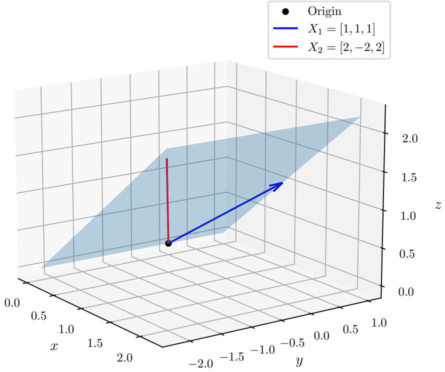
Can we obtain a point (x, y) s.t. $x = 3y$?

Can we obtain a point (x, y) s.t. $x = 3y$?
No

Can we obtain a point (x, y) s.t. $x = 3y$?
No
Span of the above set is along the line $y = 2x$




Origin
$X_1 = [1, 1, 1]$
$X_2 = [2, -2, 2]$
0.0
0.5
1.0
1.5
2.0
x
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
y
0.0
0.5
1.0
1.5
2.0
z



The span is the plane $z = x$ or $x_3 = x_1$

This condition arises when the $|X^T X| = 0$.

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} \tag{1}$$

This condition arises when the $|X^T X| = 0$.

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} \tag{1}$$

The matrix X is not full rank.

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.
It is denoted as follows {N:0, E:1, W:2, S:3 }

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.
It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding?

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.
It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding?
Then this implies that S$>$W$>$E$>$N

The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity.

The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity.
Is it S = 1 - (Is it N + Is it W + Is it E)

W and S are related by one bit.

W and S are related by one bit.
This introduces dependencies between them, and this can cause confusion in classifiers.

Encoding

Encoding

Is Female	height
1	...
1	...
1	...
0	...
0	...

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$

We get $\theta_0 = 5.9$ and $\theta_1 = -0.7$

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

height$_i$ = θ_0 + θ_1 * (Is Female) + ϵ_i

We get θ_0 = 5.9 and θ_1 = -0.7
θ_0 = Avg height of Male = 5.9

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

height$_i$ = θ_0 + θ_1 * (Is Female) + ϵ_i

We get θ_0 = 5.9 and θ_1 = -0.7
θ_0 = Avg height of Male = 5.9
$\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

height$_i$ = θ_0 + θ_1 * (Is Female) + ϵ_i

We get θ_0 = 5.9 and θ_1 = -0.7
θ_0 = Avg height of Male = 5.9
$\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).
θ_1 is chosen based on 5-5.9, 5.2-5.9, 5.4-5.9

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

height$_i$ = θ_0 + θ_1 * (Is Female) + ϵ_i

We get θ_0 = 5.9 and θ_1 = -0.7
θ_0 = Avg height of Male = 5.9
$\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).
θ_1 is chosen based on 5-5.9, 5.2-5.9, 5.4-5.9 θ_1 = Avg. female height (5+5.2+5.4)/3 - Avg. male height(5.9)

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i = \begin{cases} \theta_0 + \theta_1 + \epsilon_i & \text{if } i \text{ th person is female} \\ \theta_0 - \theta_1 + \epsilon_i & \text{if } i \text{ th person is male.} \end{cases}$$

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i = \begin{cases} \theta_0 + \theta_1 + \epsilon_i & \text{if } i \text{ th person is female} \\ \theta_0 - \theta_1 + \epsilon_i & \text{if } i \text{ th person is male.} \end{cases}$$

Now, θ_0 can be interpreted as average person height. θ_1 as the amount that female height is above average and male height is below average.

When does the normal equation have a unique solution?

When does the normal equation have a unique solution?

When does the normal equation have a unique solution?

How do polynomial features help with non-linear relationships?

When does the normal equation have a unique solution?

How do polynomial features help with non-linear relationships?

When does the normal equation have a unique solution?

How do polynomial features help with non-linear relationships?

What are the assumptions behind linear regression?

Violation Consequences:

- Biased coefficient estimates

Violation Consequences:

- Biased coefficient estimates
- Invalid confidence intervals

Violation Consequences:

- Biased coefficient estimates
- Invalid confidence intervals

Violation Consequences:

- Biased coefficient estimates
- Invalid confidence intervals
- Poor prediction performance

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix
- **Feature Engineering**: Basis expansion enables non-linear modeling

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top \mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix
- **Feature Engineering**: Basis expansion enables non-linear modeling

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix
- **Feature Engineering**: Basis expansion enables non-linear modeling
- **Foundation**: Building block for more complex models